Computational Statistics & Data Analysis 8 (1989) 227–245
North-Holland

Mixtures of linear regressions

Richard D. DE VEAUX
Princeton University, Princeton, NJ 08544, USA

Received February 1987
Revised January 1989

Abstract: The purpose of this article is to develop the technology of models based on mixtures of linear regressions and, in particular, to draw out the relevance of the EM algorithm to the associated maximum likelihood equations. A $\sqrt{n}$-consistent starting point for the EM algorithm is presented. The data from an experiment in music perception are analyzed using this technology. Performance of the estimators are examined via both simulated and actual data sets.

1. Introduction

The model of interest can be illustrated by the data of Figure 1 which come from an experiment in music perception, Cohen (1980). Even casual consideration establishes that two lines are evident. The problem is to estimate the seven parameters which specify the proportion of points on each line, the slope and intercept of each line, and the variance around each line. A mixture of linear regressions (or switching regressions) model may be appropriate if no information about membership of the points to each line is available. Here, estimators of the seven parameters are derived. Performance of these estimates are investigated via both Monte Carlo methods and application to the data shown in Figure 1.

The mixture of regressions model is given as follows:

$$y_i = \begin{cases} \alpha + \beta_1 x_i + \epsilon_{1i} & \text{with probability } \lambda \\ \alpha_2 + \beta_2 x_i + \epsilon_{2i} & \text{with probability } 1-\lambda, \end{cases} \tag{1.1}$$

where the $\epsilon_{ji} \sim N(0, \sigma_j^2)$ are independent, $j = 1, 2$, $i = 1, \ldots, n$. The seven parameters of the model can be collected conveniently in the vector $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$. We will assume throughout the paper that the design matrix X, is a $n \times 2$ matrix with the first column equal to a column of 1s and the second column equal to $(x_1, x_2, \ldots, x_n)$. We also suppose that the x_i, $1 \le i \le n$ are such that $\lim_{n\to\infty} X'X/n$ is positive definite.

Conditional on x_i, the y_i have density

$$f(y) = \lambda f_1(y) + (1-\lambda) f_2(y), \tag{1.2}$$

where $f_j(y)$ is the normal density with mean $\alpha_j + \beta_j x_i$ and variance σ_j^2, $j = 1, 2$. When $\beta_1 = \beta_2 = 0$, $f_j(y)$ is the normal density with mean μ_j and variance σ_j^2, $j = 1, 2$, where $\mu_j \equiv \alpha_j$. This five parameter ($\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$) model is the mixture of normals model which has a long history, dating back to Karl Pearson

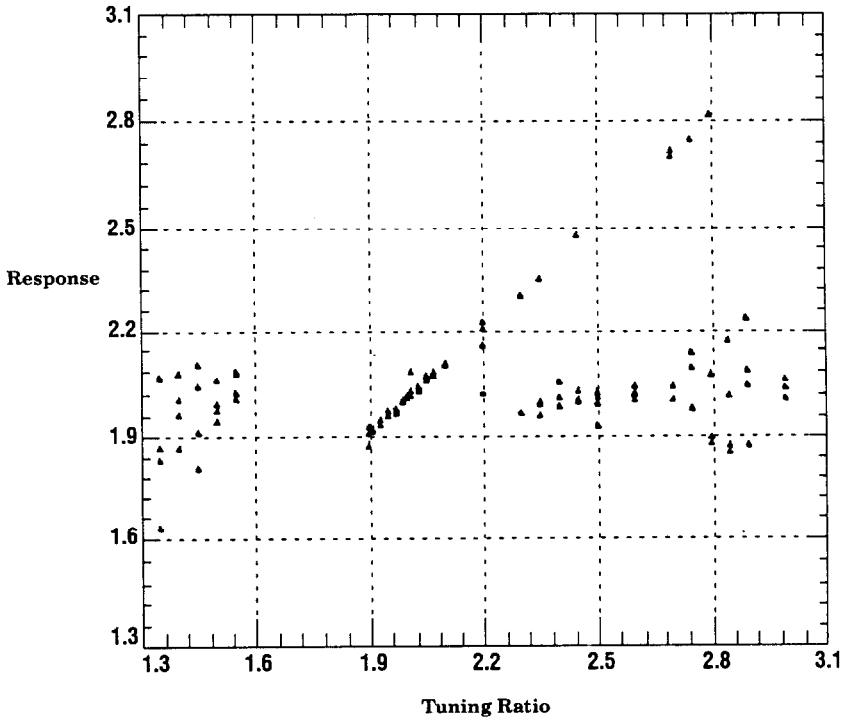

Fig. 1. Scatterplot of response vs tuning ratio.

(1894). In this case, we follow the convention that $\mu_1 \leq \mu_2$, thus distinguishing f_1 and f_2. This convention is traditional in the theory of mixture models. Analogously, we will assume that $\alpha_1 \leq \alpha_2$ for the mixture of regressions model.

Pearson developed the method of moments to obtain an estimator for the mixture of normals parameter $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. More recently, likelihood methods for the mixture of normal model have been investigated by many authors. Denote the likelihood function of a random sample $y_1, \ldots, y_n$ from a population with density (1.2) by $l(\theta) = l(y, \theta)$. Day (1969) pointed out that in the five parameter case,

$$l(\lambda, y_i, \mu_2, 0, \sigma_2^2) = l(\lambda, \mu_1, y_i, \sigma_1^2, 0) = \infty \ (i = 1, \ldots, n).$$

That is, by estimating one population to have mean y_i with arbitrarily small variance the likelihood becomes unbounded. Since the likelihood has no global maximum, Day concluded that "maximum likelihood clearly breaks down". The unboundedness of the likelihood persists in the regression case, where by estimating one line to have arbitrarily small error variance,

$$l(\lambda, y_i, \alpha_2, 0, \beta_2, 0, \sigma_2^2) = l(\lambda, \alpha_1, y_i, \beta_1, 0, \sigma_1^2, 0) = \infty.$$

In spite of this, various authors (Tan and Chang (1972), Fryer and Robinson (1972), Hosmer (1973), Dick and Bowden (1973), and Hosmer and Dick (1973)) reported good numerical convergence of likelihood maximization to local maxima. Methods not based on the likelihood have also been investigated in the literature. Quandt and Ramsey (1978) proposed minimizing the empirical mo-

ment generating function for both the five and seven parameter problems, while Woodward et al. (1984) proposed a minimum distance estimator for the five parameter case.

Kiefer (1978) showed that for the mixture of regressions problem there exists a sequence of roots of the likelihood equation which is consistent and asymptotically efficient and normally distributed. Typically such a sequence of roots are found by Newton–Raphson iterations. The closest root to the $\sqrt{n}$-consistent estimator is asymptotically efficient. Day (1969) exploited the structure of the likelihood to conditionally update the probability of membership of the data to the two component populations, obtaining iterative maximum likelihood estimators for the case of equal variances. Several authors including Hathaway (1983) and Titterington, Smith, and Makov (1985), showed that Day's estimating procedure is essentially an instance of the EM algorithm (see Dempster, Laird, and Rubin (1977) for a more complete discussion of the EM algorithm). Aitkin and Wilson (1980) used the EM algorithm in the mixture of regressions model as a means of outlier detection. The EM estimates are iteratively reweighted maximum likelihood estimates with weights equal to the current conditional probability of component membership. Explicitly, given estimates of θ at state $p-1$, weights at stage p, $W_{ji}^{(p)}$ $(j=1, 2)$, are given by

$$W_{1i}^{(p)} = \frac{\lambda^{(p-1)} f_1^{(p-1)}(y_i)}{\lambda^{(p-1)} f_1^{(p-1)}(y_i) + (1-\lambda^{(p-1)}) f_2^{(p-1)}(y_i)}, \tag{1.3}$$

where $W_{2i}^{(p)} = 1 - W_{1i}^{(p)}$. Then, the pth stage estimates of θ are given by

$$\lambda^{(p)} = \frac{\sum_{i=1}^{n} W_{1i}^{(p)}}{n}, \tag{1.4}$$

$$\alpha_j^{(p)} = \frac{\sum_{i=1}^{n} W_{ji}^{(p)} y_i}{\sum_{i=1}^{n} W_{ji}^{(p)}} - \beta_j^{(p)} \frac{\sum_{i=1}^{n} W_{ji}^{(p)} x_i}{\sum_{i=1}^{n} W_{ji}^{(p)}}, \tag{1.5}$$

$$\beta_j^{(p)} = \frac{\sum_{i=1}^{n} W_{ji}^{(p)} x_i y_i - \dfrac{\sum_{i=1}^{n} W_{ji}^{(p)} x_i \sum_{i=1}^{n} W_{ji}^{(p)} y_i}{\sum_{i=1}^{n} W_{ji}^{(p)}}}{\sum_{i=1}^{n} W_{ji}^{(p)} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} W_{ji}^{(p)} x_i\right)^2}{\sum_{i=1}^{n} W_{ji}^{(p)}}}, \quad \text{and} \tag{1.6}$$

$$\sigma_{ji}^{2(p)} = \frac{\sum_{i=1}^{n} W_{ji}^{(p)} \left(y_i - \left(\alpha_j^{(p)} + \beta_j^{(p)} x_i\right)\right)^2}{\sum_{i=1}^{n} W_{ji}^{(p)}}, \qquad j=1, 2. \tag{1.7}$$

Given starting values for the parameters, one cycles through (1.3) to (1.7) until a convergence criterion is met. Dempster, Laird and Rubin (1977) showed that at each step of the EM algorithm, the likelihood is non-decreasing: $L(\theta^{(p+1)}) \geq L(\theta^{(p)})$. Thus the EM algorithm increases to a local maximum or to the boundary, where $\sigma_j = 0$, $j = 1$ or 2. There is, however, no assurance that the EM algorithm will converge to the asymptotically efficient estimator and not to the boundary.

To avoid convergence of the EM algorithm, to $\sigma_j = 0$, $j = 1$ or 2, Hathaway (1983, 1985) considered imposing a constraint of the form

$$\min_{i,j} (\sigma_i/\sigma_j) \geq c > 0, \tag{1.8}$$

and reported good numerical convergence of the resulting constrained EM for a mixture of normals. When the EM algorithm violated the constraint, he restarted the EM at another starting value until convergence to a local maximum was achieved. He showed that the global maximizer of $L(\theta)$ under the constraint is asymptotically efficient. Thus, while a Newton-Raphson search guarantees asymptotic efficiency when started from a $\sqrt{n}$-consistent starting point, the EM finds a local maximum which may be asymptotically efficient or it converges to a singularity in the likelihood function of a boundary point.

For mixtures of univariate normals, Hathaway (1983, 1986) reported that this constrained EM was more robust to poor choices of starting values of θ and that it avoided convergence to a singularity of $L(\theta)$. However, this robustness was gained by repeatedly starting over until the constraint was not violated. In a similar study, De Veaux (1986) showed that for univariate normal mixtures, the EM when started at a $\sqrt{n}$-consistent starting point also avoided convergence to the boundary (thus inherently not violating the constraint (1.8)), and converged faster than when the EM algorithm was started from another reasonable but not $\sqrt{n}$-consistent starting point. Thus it seems that while the EM is not guaranteed to find the likelihood root closest (and hence asymptotically efficient) to the $\sqrt{n}$-consistent starting point, it does seem to do it in practice.

A $\sqrt{n}$-consistent starting point for the mixtures of regressions is presented in the next section. The performance of the resulting EM algorithm limit estimators are investigated in section 3. By comparison with the Cramér–Rao lower bounds for the variances, we find via Monte Carlo experiments that our estimators perform as well as possible. Section 4 contains a description of the experiment which motivated this investigation as well as an analysis of the experimental data. Bootstrap confidence intervals are used to test hypotheses about the parameters. We find that the parameter values proposed by Cohen (1980) are within 95% bootstrap confidence intervals. We conclude, in section 5, that the EM algorithm, when started at a $\sqrt{n}$-consistent starting point, provides estimators of the seven parameters of a mixture of regressions model which are efficient with the added features of being intuitively appealing and easy to implement.

2. An $\sqrt{n}$-consistent starting value

Pearson (1894) invented the Method of Moments (MM) to find estimators for the five parameters of a mixture of normals. To see this, for notational convenience let

$$m_1 = (1-\lambda)(\mu_1 - \mu_2) \quad \text{and} \quad m_2 = \lambda(\mu_2 - \mu_1). \tag{2.1}$$

Note that since $\mu_1 \le \mu_2$, $m_1 \le 0$ and $m_2 \ge 0$. Let α_i be the ith central moment of the mixture of normals data. Then

$$\begin{aligned}
\alpha_1 &= \lambda m_1 + (1-\lambda)m_2 = 0,\\
\alpha_2 &= \lambda\left(m_1^2 + \sigma_1^2\right) + (1-\lambda)\left(m_2^2 + \sigma_2^2\right),\\
\alpha_3 &= \lambda m_1\left(m_1^2 + 3\sigma_1^2\right) + (1-\lambda)m_2\left(m_2^2 + 3\sigma_2^2\right),\\
\alpha_4 &= \lambda\left(m_1^4 + 6m_1^2\sigma_1^2 + 3\sigma_1^4\right) + (1-\lambda)\left(m_2^4 + 6m_2^2\sigma_2^2 + 3\sigma_2^4\right),\\
\alpha_5 &= \lambda m_1\left(m_1^4 + 10m_1^2\sigma_1^2 + 15\sigma_1^4\right) + (1-\lambda)\left(m_2^4 + 10m_2^2\sigma_2^2 + 15\sigma_2^4\right).
\end{aligned} \tag{2.2}$$

Let a_i $(i = 1, \ldots, 5)$ denote the sample central moments of the data. Substitute the sample moments a_i for the α_i in the left-hand side of (2.2). We now have five equations in five unknowns $(\lambda, m_1, m_2, \sigma_1^2, \sigma_2^2)$.

After considerable algebraic manipulation, Pearson (1894), reduced the five equations of (2.2) to one nonic (ninth order) equation in one unknown (say z). The unknowns $(\lambda, m_1, m_2, \sigma_1^2, \sigma_2^2)$ are functions of z. If a negative real root of this equation can be found, say z^*, the estimates of $(\lambda, m_1, m_2, \sigma_1^2, \sigma_2^2)$ are found directly from z^* (see Pearson (1894) and Cohen (1967)).

The sample moments a_i are $\sqrt{n}$-consistent for α_i (see e.g. Serfling (1980) p. 67). Using the implicit function theorem, De Veaux (1986) showed that the MM estimator of $\hat{\theta}$ defined by (2.2) is $\sqrt{n}$-consistent for $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

Using the MM estimate above, we now proceed to derive an $\sqrt{n}$-consistent starting point for the mixture of regressions case.

1. Choose two non-overlapping intervals (or bins) on the x-axis, each containing 100 p% of the x_i. In each bin $I^{(k)}$, $k = 1, 2$, consider only the y_i. The y_i are approximately a mixture of normals, with $\theta^{(k)} = (\lambda, \mu_1^{(k)}, \mu_2^{(k)}, \sigma_1^{2(k)}, \sigma_2^{2(k)})$. (They would be precisely a mixture of normals if they all had the same x_i value.) Use the MM estimate to estimate $\theta^{(k)}$ in each bin. We display the data as $(\bar{x}^{(k)}, y_i)$ where $\bar{x}^{(k)}$ is the average of x_i in bin k, in Figure 2. Also indicated are the MM estimates $\hat{\mu}_1^{(1)}$, $\hat{\mu}_1^{(2)}$, $\hat{\mu}_2^{(1)}$, and $\hat{\mu}_2^{(2)}$.
2. We use the four MM estimates $(\hat{\mu}_1^{(1)}, \hat{\mu}_2^{(1)}, \hat{\mu}_1^{(2)}, \hat{\mu}_2^{(2)})$ to construct two possible pairs of lines. One pair connects the upper means, $\hat{\mu}_2^{(1)}$ and $\hat{\mu}_2^{(2)}$ to each other, and the lower means, $\hat{\mu}_1^{(1)}$ and $\hat{\mu}_1^{(2)}$ to each other. The lines formed by this do not intersect between $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ and these are referred to as the straight pair of lines (see Figure 3a). Alternatively, we form a pair of lines connecting $\hat{\mu}_1^{(1)}$ to $\hat{\mu}_2^{(2)}$ and $\hat{\mu}_2^{(1)}$ to $\hat{\mu}_1^{(2)}$. This pair does intersect in $(\bar{x}^{(1)}, \bar{x}^{(2)})$ and is called the crossed pair (see Figure 3b). Each pair of lines defines two intercept and slopes which are denoted $(\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s})$ and $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c})$ respectively.

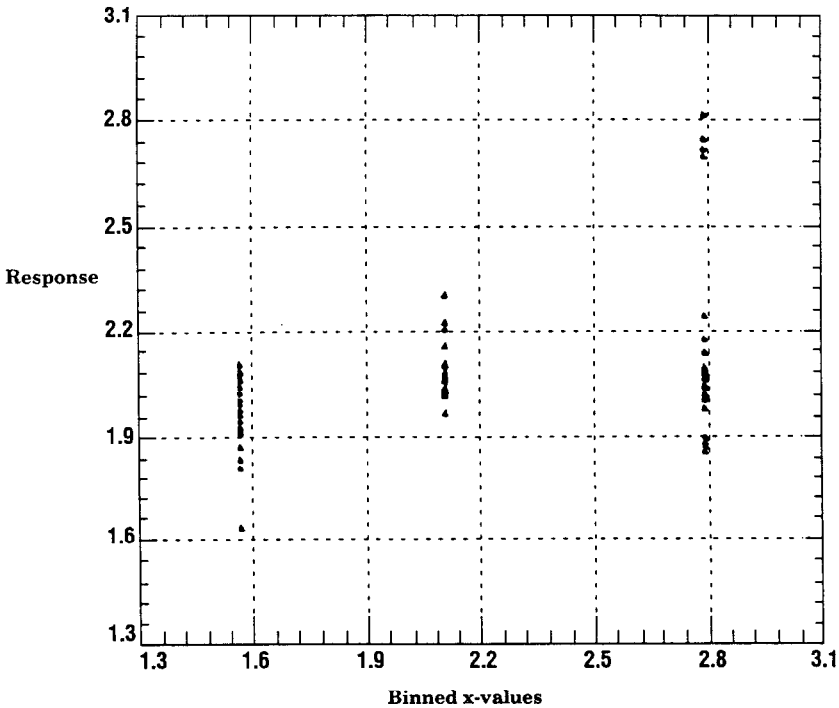

Fig. 2. Plot of binned data.

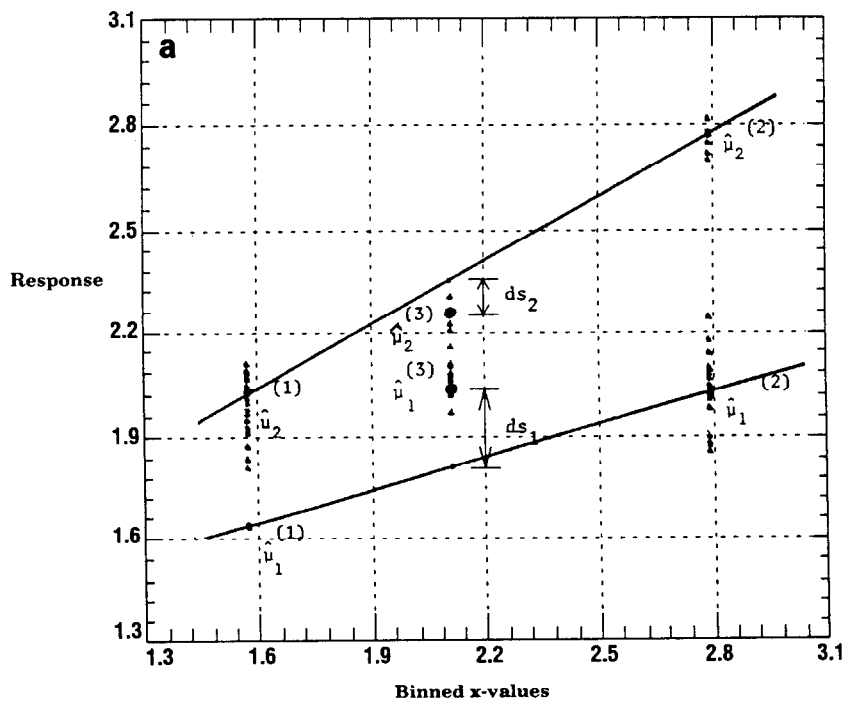

Fig. 3a. Plot of binned data with straight lines.

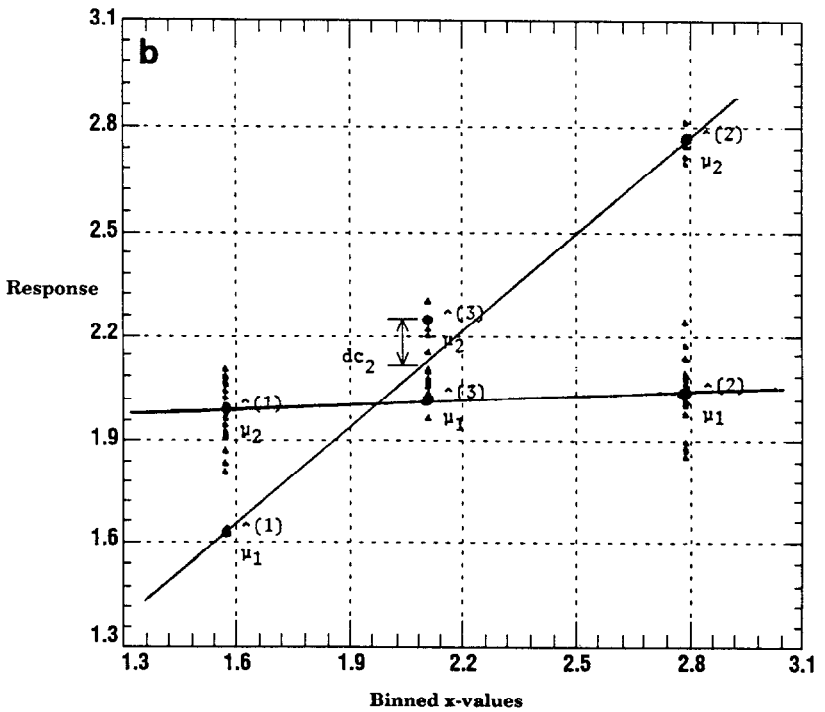

Fig. 3b. Plot of binned data with crossed lines.

3. One set of intercept and slope estimates are $\sqrt{n}$-consistent for $(\alpha_1, \alpha_2, \beta_2, \beta_2)$ (see De Veaux (1986)). We use a third bin to choose between them. Let $I^{(3)}$ denote such a third non-overlapping interval on the x-axis also containing 100 p% of the x_i. Follow the above procedure and estimate $\hat{\mu}_1^{(3)}$ and $\hat{\mu}_2^{(3)}$ for the y_i in $I^{(3)}$ (see Figure 2). We will choose the pair of lines with smaller distance to $\hat{\mu}_1^{(3)}$ and $\hat{\mu}_2^{(3)}$ (see Figures 3a and 3b). Specifically, calculate the absolute distance from the upper mean estimate $\hat{\mu}_2^{(3)}$ to the upper line at $\bar{x}^{(3)}$, and denote it by d_{s_2}. That is,

$$d_{s_2} = \left|\hat{\mu}_2^{(3)} - \left(\hat{\alpha}_{2s} + \hat{\beta}_{2s}\bar{x}^{(3)}\right)\right|.$$

Similarly, calculate the absolute distance from $\hat{\mu}_1^{(3)}$ to the lower line and denote it by d_{s_1}. We define the total error of this pair as $d_s = d_{s_1} + d_{s_2}$. We repeat the analogous procedure for the crossed pair, and calculate $d_c = d_{c_1} + d_{c_2}$. We choose the pair with the smaller total error.

4. We now need starting points for λ and the error variances σ_j^2, $j = 1, 2$. Consider any of the three bins. For the appropriate line, use $\hat{\lambda}$ from the MM estimate. The variances of the binned y_i as estimated by the MM will in general be larger than the σ_j^2 since x_i varies within the bin. We adjust the variance estimates as follows: in each bin, obtain the EM estimates of the five parameter mixture from the MM estimate starting points. The EM algorithm

variance estimates will be of the form

$$\hat{\sigma}_j^{*2} = \frac{\sum W_{ji}^* (y_i - \mu_j^*)^2}{\sum W_{ji}^*}, \tag{2.3}$$

$j = 1, 2$. The * indicates that these are the limit estimators provided by the EM algorithm limit. This variance estimate is too large because the term μ_j^* in (2.3) is constant over the interval whereas the true conditional means of each y_i in bin k are $\alpha_j + \beta_j x_i$ $j = 1, 2$ and not $\mu_j^{(k)} = \alpha_j + \beta_j \bar{x}^{(k)}$. Thus, we substitute the estimates $\hat{\alpha}_j + \hat{\beta}_j x_i$ for μ_j^* in (2.3), and obtain

$$\hat{\sigma}_j^2 = \frac{\sum W_{ji}^* \left(y_i - \left(\hat{\alpha}_j + \hat{\beta}_j x_i \right) \right)^2}{\sum W_{ji}^*}. \tag{2.4}$$

In summary, we have the following proposition:

Proposition. The estimator $\hat{\theta}_n = (\hat{\lambda}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_1^1, \hat{\sigma}_2^2)$ *described in* 1-4 *above is* $\sqrt{n}$*-consistent for* θ. (*Details of the proof can be found in De Veaux* (1986), *Lemmas* 3.31–3.35.)

Some practical considerations should be pointed out. Method of moment estimators for mixtures of univariate normals perform best when the overlap of the univariate normals is small. Thus choosing a bin on the x-axis which has a small overlap in the y_i will result in more stable starting points for the procedure. If a crossing point is evident it is best to avoid it in choosing the bins. The estimate of λ is $\sqrt{n}$-consistent for any bin. However, averaging the estimates obtained from the three bins results in a more stable starting point.

Our $\sqrt{n}$-consistent estimator

$$\hat{\theta}_n = (\hat{\lambda}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2) \tag{2.5}$$

is then the starting point for the EM algorithm for the seven parameter problem. That is, we let $\theta^{(p)}$ be the pth step of the EM algorithm with $\theta^{(0)} = \hat{\theta}_n$. Define the resulting EM algorithm limit estimator by $\theta^* = \lim_{p \to \infty} \{\theta^{(p)}\}$. The performance of θ^* will be discussed in section 3.

3. Monte Carlo simulation

A Monte Carlo simulation was performed to see if the procedure outlined in the previous section provides reasonable estimates in practice. The scope was limited to the study of three cases. The first case was designed to resemble the data from the musical perception experiment. Specifically we set $\theta = (0.5, 0, 2, 1, 0, 0.03^2, 0.05^2)$ with x_i ranging from 1.5 to 3.0. For comparison we also investigated two higher overlap situations, $\theta = (0.33, 0, 1, 1, -1, 0.1^2, 0.2^2)$, and $\theta = (0.5, 0, 1, 1, -1, 0.2^2, 0.3^2)$, with x ranging from 0 to 1. We refer to these as configuration 1, 2, and 3 respectively. Scatterplots of random sample of 100

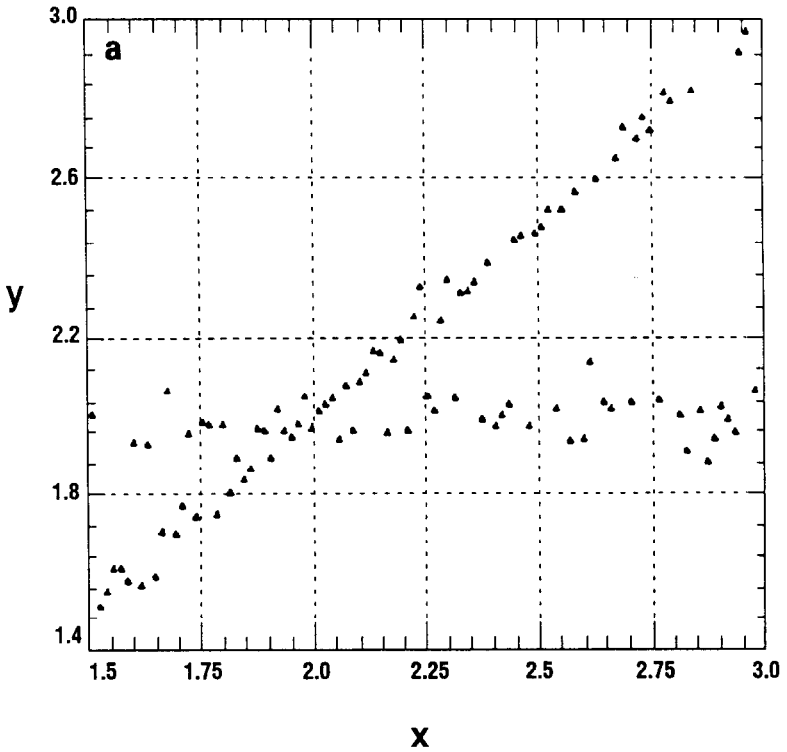

Fig. 4a. Scatterplot of sample from configuration 1.

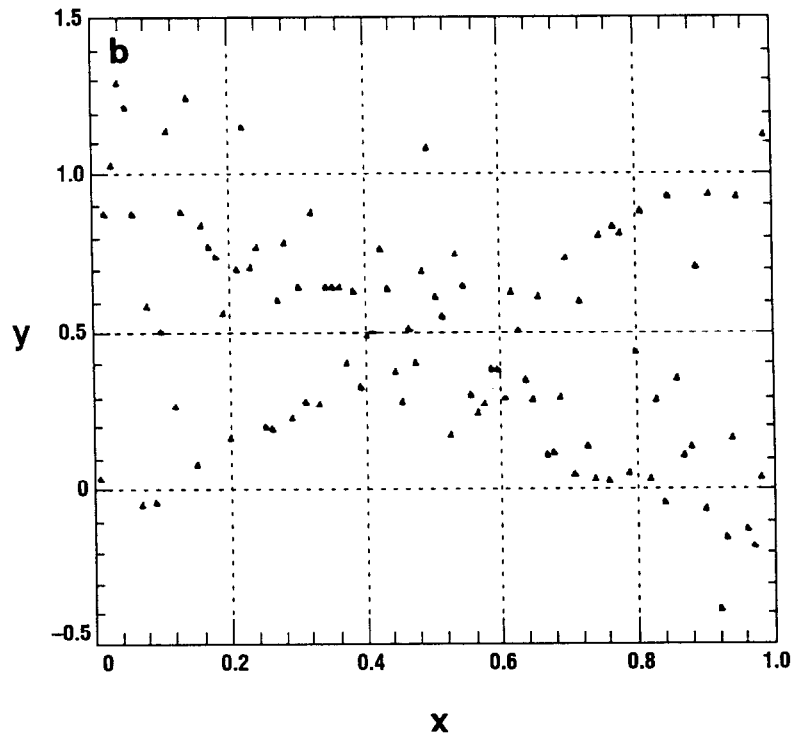

Fig. 4b. Scatterplot of sample from configuration 2.

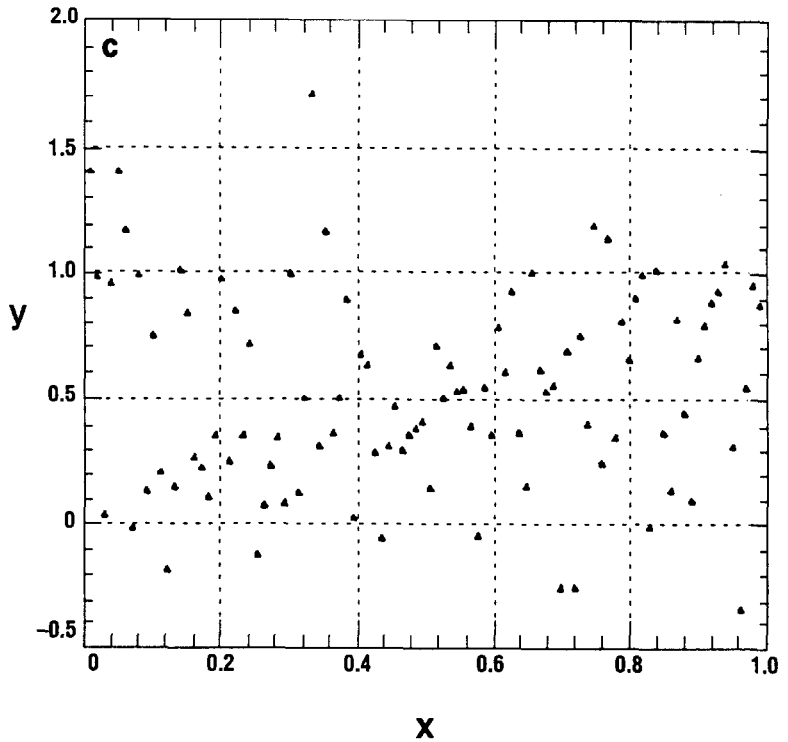

Fig. 4c. Scatterplot of sample from configuration 3.

points from each of these models are shown in Figures 4a, 4b, and 4c. Note that in all three situations, the true regression lines do cross within the range of the x variable. For this simulation, programs were written in APL and run on an IBM-PC AT with 80287 numeric co-processor. At each x_i a Bernoulli (λ) random variable was generated to assign y_i to the regression lines. The random number generators were based on the uniform random generator supplied by STSC^{TM} APL version 2.0. Further details can be found in De Veaux (1986).

A first concern is how well the binning estimates did in choosing the correct (i.e., the crossed) pair of lines as starting points, and how well the EM algorithm subsequently did in finding the correct pair when it did not. For each configuration, 100 simulations of 100 pairs (x_i, y_i) were generated according to (1.1), with the x_i evenly spaced in their range. The binning starting points chose a crossed starting point 91%, 92% and 63% in the three configurations. In all 300 cases however, the EM algorithm limits were a pair of crossed lines. While the binning procedure does not always choose the correct starting lines, the EM algorithm did (using a convergence criterion of $\epsilon = 10^{-6}$ in the sample likelihood function), and in particular never converged to the boundary at σ_1 or $\sigma_2 = 0$. Thus, the $\sqrt{n}$-consistent starting point precluded the use of Hathaway's (1985) constraint (1.8) on σ_1/σ_2. Only one starting point was used.

For each configuration, the information matrix was obtained via Romberg's algorithm (see Behboodian (1972)), and from this the Cramér–Rao lower bound

for the standard errors were calculated. We also calculated the standard errors of the estimates had the populations memberships been known. We refer to these standard errors as the full information standard errors. We will use both of these to compare with the sample standard errors.

Configuration 1: $\theta = (0.5, 0, 2, 1, 0, 0.03^2, 0.05^2)$ For each parameter estimate, the average, the standard errors from the information matrix, the sample covariance matrix and full information are:

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
True	0.500	2.000	0.000	0.000	1.000	0.0500	0.030
Average of 100 runs	0.501	1.999	0.000	0.001	0.995	0.049	0.028
Sample (std. dev.)	0.0548	0.0414	0.0247	0.0174	0.0106	0.0207	0.0143
Information (std. dev.)	0.0500	0.0225	0.0376	0.0098	0.0163	0.0134	0.0223
Full Information (std. dev.)		0.0222	0.0375	0.0098	0.0163		

For this configuration, we see that the information matrix standard errors are nearly the same as the full information standard errors. This is due to the extremely low overlap of the configuration. The only curious feature of the comparison is that for α_2, β_2, and σ_2, the simulation standard errors are significantly ($\alpha = 0.05$) less than the information matrix standard errors. This may be due to the sample correlation of the estimates which in general are more correlated than predicted by the information matrix. This behavior was not seen in the next two configurations.

Configuration 2: $\theta = (0.33, 0, 1, 1, -1, 0.1^2, 0.2^2)$:

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
True	0.333	0.000	1.000	1.000	−1.000	0.100	0.200
Average of 100 runs	0.329	0.001	0.999	0.997	−0.996	0.094	0.193
Sample (std. dev.)	0.0588	0.0414	0.0476	0.0756	0.0845	0.0456	0.0930
Information (std. dev.)	0.0471	0.0353	0.0494	0.0607	0.0850	0.0503	0.0840
Full Information (std. dev.)		0.0350	0.0486	0.0601	0.0841		

For this configuration, again a relatively low overlap configuration, the full information standard errors are comparable to the information matrix standard errors. The sample standard errors are in general slightly larger than predicted, but none are significantly so (overall $\alpha = 0.05$).

Configuration 3: $\theta = (0.5, 0, 1, 1, -1, 0.2^2, 0.3^2)$:

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
True	0.500	0.000	1.000	1.000	−1.000	0.200	0.300
Average of 100 runs	0.506	−0.019	0.990	1.008	−0.991	0.193	0.291
Sample (std. dev.)	0.1104	0.0829	0.1479	0.1511	0.2677	0.1229	0.1658
Information (std. dev.)	0.0565	0.0640	0.981	0.1069	0.1628	0.1024	0.1658
Full Information (std. dev.)		0.0559	0.0874	0.0980	0.1473		

For this configuration, we see that the asymptotic standard errors are about 10% higher than the full information standard errors, reflecting the comparatively high overlap of this configuration. Five of the sample standard errors are significantly larger than the information matrix standard errors the exceptions being α_1 and σ_2 (overall $\alpha = 0.05$).

The binding procedure, as expected, chooses the correct pair of starting lines with increasing frequency as the overlap decreases. The EM algorithm correctly chose a crossed line solution in all 300 simulations even when the binning procedure started at a straight line starting solution. The $\sqrt{n}$-consistent starting point seemed to point the EM algorithm to the correct root of the likelihood equation. Hathaway's constraint (1.8) was never violated. We should reiterate that it is, however, always possible for the EM to converge to a spurious maximum if started far enough away from the true parameters.

4. Data analysis

This investigation was motivated by the data from an experiment in musical perception carried out at the Center for Research in Music and Acoustics (CCRMA) at Stanford University by E.A. Cohen (Cohen (1980)). With the advent of electronically produced tones, a composer has the option to change the relationship of the overtones to the fundamental frequency of a tone. Traditionally, in Western definite pitched instruments, the overtones are at integer multiples of the fundamental. The purpose of the experiment was to see how changing the ratio of the overtones to the fundamental affects the perception of the tone.

Specifically, subjects (trained musicians) are presented with a pure fundamental tone plus a series fo overtones which are stretched (or compressed) logarithmically. (The nth overtone $f_n = (2/x)^8 x^{\log_2 n f_1}$). The factor x is called the stretching ratio. A stretching ratio of 2.0 corresponds to the harmonic pattern usually heard in traditional definite pitched instruments. The subjects are asked to tune an adjustable tone to the octave above the fundamental tone. One theory of musical perception, the interval memory hypothesis, predicted that the subjects would tune the tones to the (nominal) octave at ratio 2 : 1 to the fundamental frequency, regardless of the stretching factors. An alternative, the partial matching hypothesis, predicted that the subject would use the overtones to tune the tone and thus tune it to x, the stretching ratio. The experiment was designed to determine which theory, if either, was tenable.

The data for each of five subjects consists of pairs (x_i, y_i) where x_i is the stretching ratio of the ith trial and the response y_i is the ratio of the adjusted tone to the fundamental. The stretching ratio ranged from 1.35 to 3.0, and the number of trials for each subject ranged from 112 to 360. The data from one subject are displayed in Figure 1. (The experiment is described in full detail in Cohen (1980 and 1984).)

From Figure 1, two lines are evident: a line fo slope 1 through the origin and a line of slope 0 at $y = 2.0$. These two lines correspond to the behavior predicted by

the two theories. The following model was hypothesized by Cohen (1980 and 1984) in her data analysis:

$$y_i = \begin{cases} x_i + \epsilon_{1i} & \text{with probability } \lambda \\ 2.0 + \epsilon_{2i} & \text{with probability } 1-\lambda, \end{cases} \tag{4.1}$$

where the $\epsilon_{ij} \sim N(0, \sigma_j^2)$ $j = 1, 2$ are independent. This is seen to be a special case of model (1.1) with $\alpha_1 = 0$, $\alpha_2 = 2$, $\beta_1 = 1$ and $\beta_2 = 0$.

We select three bins, each containing 24% of the data: $I^{(1)} = [1.35 \text{ to } 1.91]$, $I^{(2)} = [2.60 \text{ to } 3.00]$ and $I^{(3)} = [2.01 \text{ to } 2.30]$. In each bin, we replace the pair (x_i, y_i) by $(\bar{x}, y_i)$ where $\bar{x}$ is the average of the x-values in the bin. This is illustrated for this data set in Figure 2.

The estimates of the two means in each bin are shown in Figure 2. These are denoted by $\hat{\mu}_j^{(k)}$ for $j = 1, 2$ and $k = 1, 2, 3$. The estimates of λ from bin 1, bin 2 and bin 3 are 0.04, 0.61 and 0.61 respectively. The extremely low estimate from bin 1 is due to the single point which appears below the rest in bin 1. The implications of this will be discussed at the end of this section.

From the estimates of the means in the outer bins, we calculate the slopes and intercepts of the straight and crossed pairs of lines, and find $(\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s}) = (0.97, 1.15, 0.63, 0.31)$ and $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c}) = (0.22, 1.90, 0.90, 0.04)$. See Figures 3a and 3b.

We must now choose between these two pairs. To do this, we use the mean estimates from bin 3, the middle bin. Start with the straight pair of lines. Refer to Figure 3a, and note that at $\bar{x}^{(3)} = 2.1$, the lower line underestimates the lower mean of the y-values in bin 3 by $ds_1 = 0.084$, while the upper line overestimates the upper mean of the y-values in bin 3 by $ds_2 = 0.209$. Thus $d_s = 0.293$. We similarly calculate the distance to the crossed pair and find $d_c = 0.204$. Thus, we choose the crossed pair as the starting point, and use $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c}) = (0.22, 1.90, 0.90, 0.04)$. The initial estimate for λ is obtained by averaging λ in all three bins resulting in $\hat{\lambda} = 0.42$. The initial standard deviation estimates are obtained via equation (1.6) and equal 0.010 and 0.12. Thus, the starting value $\hat{\theta}_n$ is $(0.42, 0.22, 1.90, 0.90, 0.04, 0.010^2, 0.12^2)$.

This estimate is now used as the starting value for the EM algorithm. The EM algorithm limit obtained was:

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
$\theta^* =$	(0.403,	−0.005,	1.90,	1.000,	0.040	0.0070,	0.090)

compared with the starting point

$$\hat{\theta}_n = \quad (0.420, \quad 0.220, \quad 1.90, \quad 0.900, \quad 0.040, \quad 0.0100, \quad 0.120).$$

The data with the EM lines superimposed are shown in Figure 5.

The EM algorithm estimator does not assign the points to each line, but rather for each point y_i, it assigns probabilities W_{1i}^* and W_{2i}^* $(W_{1i}^* + W_{2i}^* = 1)$ of belonging to the first or second line of (1.1). (See (1.7)). Because of this, we can not calculate residuals as in an ordinary linear regression situation. (See also Cox

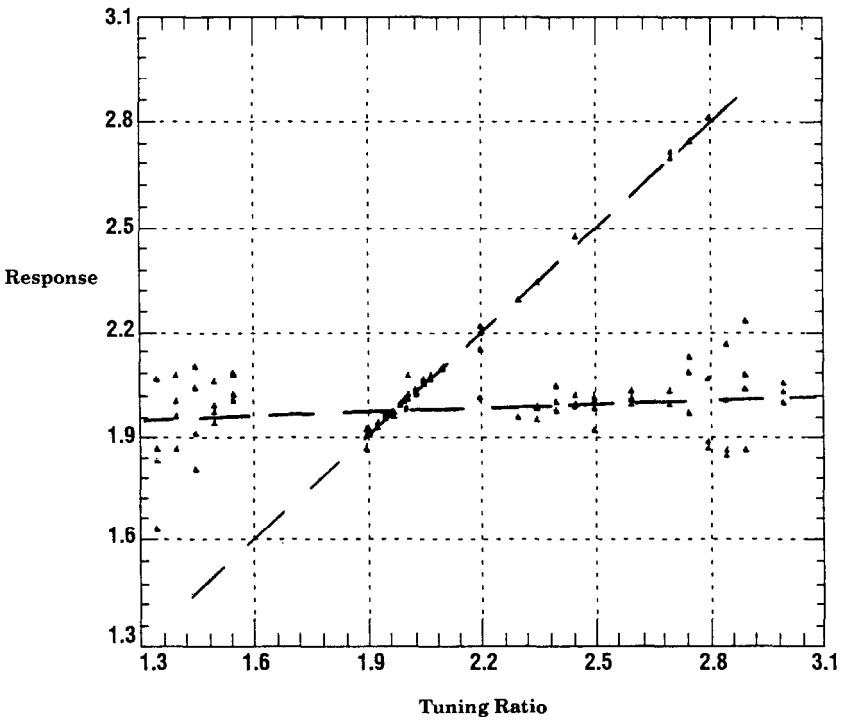

Fig. 5. Plot of data with EM lines.

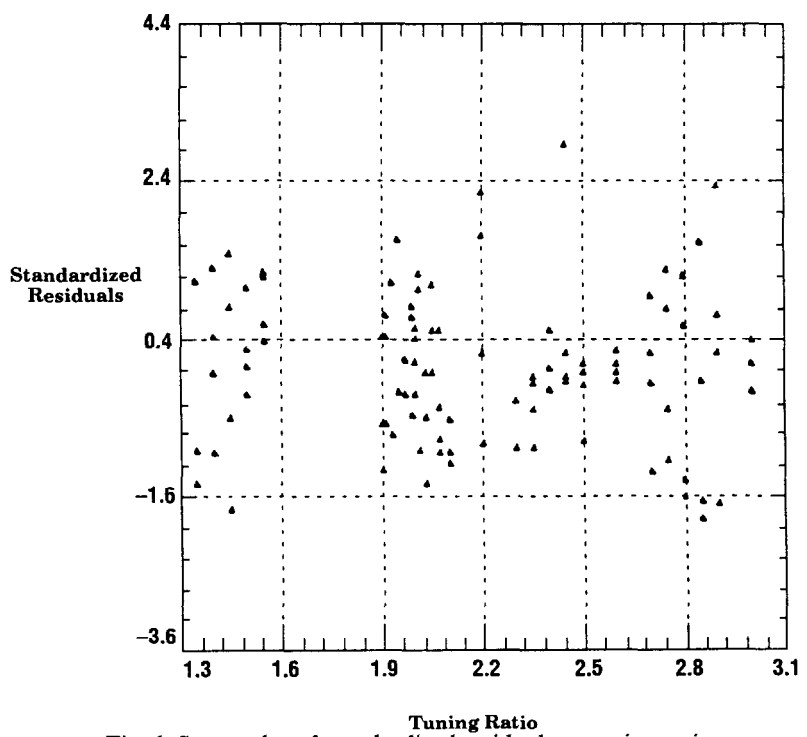

Fig. 6. Scatterplot of standardized residuals vs tuning ratio.

and Snell (1968)). Rather, we shall define residuals in the following way. We assign points to each line on the basis of whether the final EM algorithm probability W_{1i}^* is greater or less than 0.5. The residuals are then calculated in the usual way, $r_i = y_i - (\hat{\alpha}_j + \hat{\beta}_j x_i)$ where j is 1 or 2 if W_{1i}^* is $>$ or < 0.5 respectively. The r_i will be approximately a mixture of normals each with mean 0, since the true errors ϵ_{ij} are a mixture of $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$. The standardized residuals, e_i are defined to be $e_i = r_i/\hat{\sigma}_j$. Figure 6 is a plot of the e_i versus x_i. No obvious patterns remain. The standardized residuals should be approximately normally distributed. (The true residuals if population membership were known would be normally distributed). The only standardized residual greater than 2.33 in absolute value is $e_1 = -3.589$. This point was removed to see the influence on the EM estimates. The removal of the point resulted in a 10% decrease in the standard deviation estimate of the line with estimated slope 0.04 and intercept 1.90. The other six parameters were practically unchanged. Because no scientific justification could be made for deleting this point, it was left in the subsequent analysis.

Confidence intervals are obtained from the standard errors of bootstrap replications of the data. (See Efron (1977) for an introduction to the bootstrap). The data (x_i, y_i) are resampled with replacement and θ is re-estimated by the above procedure at each resampling (replication). The standard errors of the parameter estimates of these replications then provide estimates of the actual standard errors of the estimators.

The data (x_i, y_i) from the subject were resampled with replacement 500 times. The EM estimate was obtained for each replication. The average values of $\hat{\theta}$ are shown with the original data EM estimates for comparison:

Original EM estimate

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
$\hat{\theta} =$	(0.403,	−0.00462,	1.902,	1.003,	0.0402,	0.00716,	0.0901).

Average of 500 Bootstrapped EM estimates:

$\hat{\theta} =$	(0.405,	−0.00429,	1.900,	1.003,	0.0146,	0.00679,	0.0880)
	(0.053)	(0.0144)	(0.060)	(0.0071)	(0.025)	(0.0036)	(0.0431)

where the bootstrap sample standard errors of these 500 replications appear in parentheses below each parameter. For comparison, the information matrix was calculated for these x_i at $\theta = \theta^*$, the EM estimate. The bootstrap standard errors and the information matrix standard errors are:

	λ	α_1	α_2	β_1	β_2	σ_1	σ_2
Bootstrap	0.053	0.0144	0.060	0.0071	0.025	0.0036	0.0431
Information	0.046	0.0051	0.053	0.0023	0.024	0.0033	0.0375

Thus we see that the information matrix standard errors are smaller than the bootstrap with the largest discrepancies occurring for α_1 and β_1. Here the ratios of the bootstrap to the information matrix standard errors are 3.54 and 3.24. This may be due to a few data points which are high influence points for determining the slope and intercept of the slope 1 line. Because the standard deviation around this line is so small, in those replications where the high influence points are selected or even repeated, the change in the estimates of α_1 and β_1 will be higher than predicted from the information matrix estimates which assume that all of the assumptions of the mixture model are true. We use the bootstrap standard errors to derive approximate confidence intervals for the parameters by

$$\theta_i \in (\hat{\theta}_i \pm 1.96 s_i^*),$$

where s_i^* is the bootstrap standard error. For the subject, the 95% bootstrap confidence intervals are:

λ	α_1	α_2	β_1	β_2	σ_1	σ_2
0.300	−0.033	1.79	0.99	−0.00964	0.00516	0.0692
0.507	0.024	2.02	1.02	0.0900	0.00917	0.111

This now enables us to test the hypothesis that the data are from a mixture of the two perception paradigms postulated in the introduction. That is, we test whether the mixture is of the form (4.1) with $\alpha_1 = 0.0$, $\beta_1 = 1.0$ and $\alpha_2 = 2.0$, $\beta_2 = 0.0$. From the confidence intervals, we see that indeed, all four postulated values are within the confidence limits. Notice also that $\sigma_2 \approx 10\sigma_1$ indicating that when the subjects used the partials to tune the octave, they did so with much greater precision.

The final assumption to be considered is whether λ is constant over the range of x. To check this, we have used the weights W_{1i}^*, the EM probabilities of belonging to the first line as indicators for λ. If we plot W_1^* versus x and y (Figure 7), we see that most of the points with high values of W_1^* (high probability fo membership to the slope 1 line) have x values between 1.90 and 2.10. Below $x = 1.90$, there are 20 points, all of which have W_{1i}^* near 0, indicating that all of these points are from the "flat line" of intercept 2, slope 0. Similarly, for x above 2.10, 39 of the 49 points have W_{1i}^* near 0. To graphically illustrate this, we have plotted (Figure 8) a running average of $\hat{\lambda}$ versus x and y. (Thus, the first point is $(x_1, y_1, W_{11}^*) = (1.35, 1.60, 0)$ while the last is $(x_{112}, y_{112}, \Sigma_{i=1}^{112} W_{11}^*/112) = (3.00, 2.04, \hat{\lambda})$). It is graphically clear that λ is not constant over the x_i. This has profound implications for the scientific question at hand. Apparently, for this subject, the ability to hear the octave (represented by the points on the flat line), increases for tuning ratios (x_i) which lies farther away from $x_i = 2.00$ in both directions. One could run separate analyses, and obtain separate confidence intervals for λ from these three regions to test this, but three things seem clear from our present analysis. First, from our exploratory analysis, the assumption of constant λ is clearly violated. Secondly, the binning estimates will suffer from non-constant λ, since for the first third of the data, nearly all the

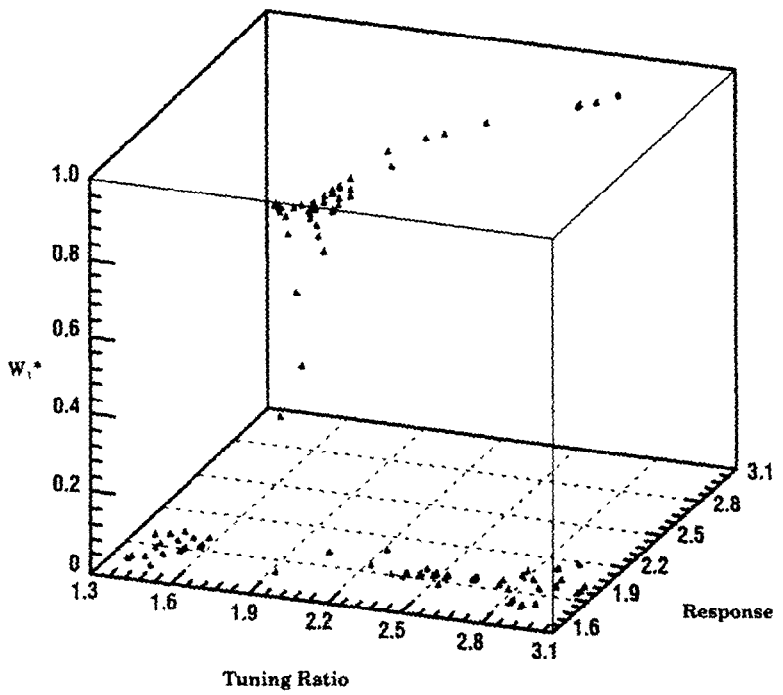

Fig. 7. Plot of W_1^* vs tuning ratio and response.

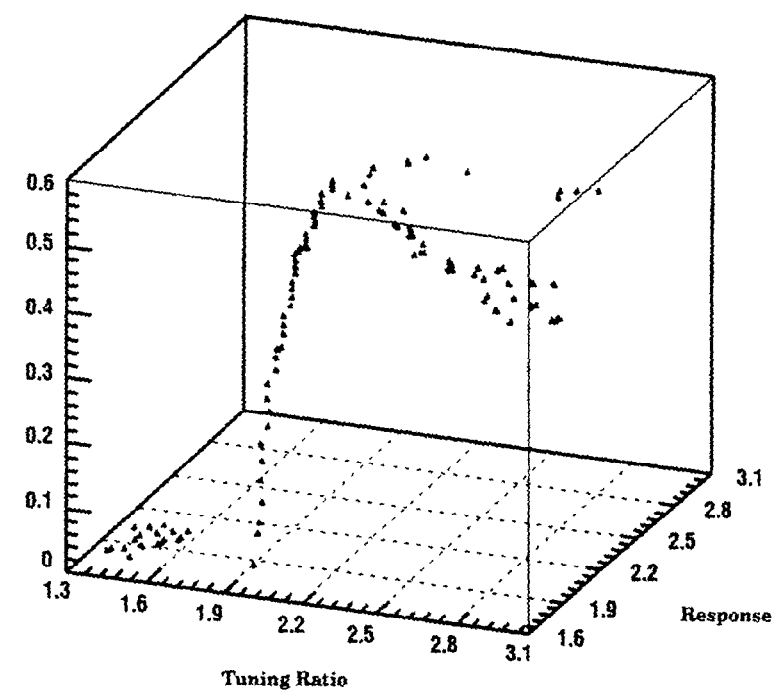

Fig. 8. Plot of $\hat{\lambda}$ vs tuning ratio and response.

y_i are from the same line. The estimate of λ from the first bin was 0.04 while from the other two it was 0.61 due to varying λ. For this data set, however, both the binning starting value and the resulting EM estimate appear to be unaffected by this problem.

The EM estimator seems to be robust to violations of this assumption since there appears to be no pattern in the residuals indicating that this problem caused the EM algorithm to allocate points to the wrong line. On the contrary, it seemed able to estimate the overall λ probability despite the fact that λ was not constant over the range of x.

5. Conclusions

We conclude that the parameters of the model hypothesized by Cohen (1980) are within the 95% bootstrap confidence limits and thus that the data can be viewed as a mixture of two lines, one with intercept 0 and slope 1, the other with intercept 2 and slope 0. The assumption of constant λ over the range of x was tested graphically and does not seem to hold up. However, if the data are viewed as having constant λ, this overall λ seems to be estimated accurately by the procedure. A further step in the data analysis might be to model λ as a function of x, or to perform separate analyses for different ranges of x.

We have provided a $\sqrt{n}$-consistent starting point for the EM algorithm, and shown that this results in estimators of the seven parameters of the mixture of regressions model (1.1). We have demonstrated their performances via three simulated data sets and on an actual data set that the EM estimates had small bias and were within the accuracy predicted by asymptotic efficiency.

Acknowledgements

This work comprised part of the author's doctoral dissertation at Stanford University. The author would like to thank Persi W. Diaconis for serving as his advisor. The author would also like to thank J. Michael Steele and an anonymous referee for comments on an earlier draft of this paper, and Elizabeth A. Cohen for providing the data that motivated the research.

References

Aitkin, M. and Wilson, G.T. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics* **22**, 325–331.

Behboodian, J. (1972). Information matrix for a mixture of two normal distributions, *Journal of Statistical Computation and Simulation* **1**, 295–314.

Cohen, A.C. (1967), Estimation in mixtures of two normal distributions, *Technometrics* **9**, 15–28.

Cohen, E.A. (1980). Inharmonic tone perception, Unpublished Ph.D. Dissertation, Stanford University.

Cohen, E.A. (1984). Some effects of inharmonic partials on interval perception, *Music Perception* **1**, 323–349.

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society B* **30**, 248–275.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions, *Biometrika* **56**, 463–474.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B* **39**, 1–38.

De Veaux, R.D. (1986). Parameter estimation for a mixture of linear regressions, Technical Report No. 247, and unpublished Ph.D. dissertation, Statistics Department, Stanford University.

Dick, N.P. and Bowden, D.C. (1973). Maximum likelihood estimation of two normal distributions, *Biometrics* **29**, 781–780.

Efron, B. (1977). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.

Fryer, J.G. and Robertson, C.A. (1972). A comparison of some methods for estimating mixed normal distributions, *Biometrika* **59**, 639–648.

Hathaway, R.J. (1983). Constrained maximum likelihood estimation for a mixture of multivariate normal densities, Technical Report 92, Dept. Math. Stat., University of South Carolina, Columbia, S.C.

Hathaway, R.J. (1985). A constrained formulation of maximum likelihood estimation for normal mixture distributions, *Annals of Statistics* **13**, 795–800.

Hosmer, D.W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of two normal distributions under three different types of samples, *Biometrika* **29**, 761–770.

Hosmer, D.W. and Dick, N.P. (1973). Maximum likelihood estimation for mixtures of two normal distributions, *Biometrics* **29**, 781–790.

Kiefer, N.M. (1978). Discrete parameter variation: efficient estimation of a switching regression model, *Econometrics* **46**, 427–434.

Pearson, K. (1894). Contribution to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society A* **185**, 71–110.

Quandt, R.E. and Ramsey, J.B. (1978).l Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association* **73**, 730–738.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Tan, W.Y. and Chang, W.C. (1972) Some comparisons for the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities, *Journal of the American Statistical Association* **67**, 792–808.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, New York.

Woodward,W.A., Parr, W.C., Schucany, W.R., and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion, *Journal of the American Statistical Association* **79**, 590–598.